DOCUMENT RESUME

ED 448 736 IR 020 499

AUTHOR Torok, Andrew G.
TITLE Indexing and Metatag Schemes for Web-Based Information Retrieval.
PUB DATE 1999-10-00
NOTE 6p.; In: WebNet 99 World Conference on the WWW and Internet Proceedings (Honolulu, Hawaii, October 24-30, 1999); see IR 020 454.
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Computer Uses in Education; Educational Technology; Evaluation Criteria; Higher Education; Indexes; *Indexing; *Information Retrieval; *World Wide Web
IDENTIFIERS Dublin Core; *Metadata; Northern Illinois University; *Web Based Instruction; XML

ABSTRACT

This paper reviews indexing theory and suggests that information retrieval can be significantly improved by applying basic indexing criteria. Indexing practices are described, including the three main types of indexes: pre-coordinate, post-coordinate, and variants of both. Design features of indexes are summarized, including accuracy, consistency, exhaustivity, and specificity. World Wide Web-based indexing is addressed, focusing on the use of meta tags. Two meta tag schemes currently in use, XML (eXtensible Markup Language) and the Dublin Core Metadata set, are discussed as a means for implementing indexing theory for Web-based documents. A project using the Dublin Core in support of a survey class in educational practice and pedagogy at Northern Illinois University is presented. (Contains 14 references.) (MES)

Indexing and Metatag Schemes for Web-based Information Retrieval

Dr. Andrew G. Torok
Northern Illinois University, Department of Educational Technology, Research and Assessment
Tel: (815) 753-3406, Fax: (815) 753-9388, E-mail: atorok@niu.edu

Abstract: Web-based searches typically result in lower precision than with other document retrieval systems. A major contributor to low precision is limited technology for generating detailed document and content descriptions such as those associated with bibliographic databases. The result is ineffective indexing of Web pages. This paper reviews indexing theory and suggests that authors can contribute significantly to improving retrieval by applying basic indexing criteria. The Dublin Core Meta Data set and XML are seen as a means for implementing indexing theory for web-based documents. A project using the Dublin Core in support of a university class is presented.

1. Introduction

Ideally, when searching for information, one should have access to the world's information resources. Traditional indexing and classification systems have done a reasonably good job of providing access to the bibliographic universe. In today's world, convenience dictates use of information retrieval tools. As a result, traditional retrieval systems are often overlooked in preference for the World Wide Web. Estimates indicate that over 95 percent of all published materials are not accessible through the Web. In addition to not having access to the bibliographic universe, Web searches result in significantly lower Precision and Recall. [Torok & West, 1999] Precision is a standard measure that provides a percentage estimate of the number of relevant items retrieved out of all items retrieved. Recall is the percentage estimate of all relevant items retrieved out of all relevant items in a database. These measures reflect the relevance judgments associated with document retrieval. Although a number of factors contribute to low Web-based retrieval performance, the most significant ones are those resulting from inadequate indexing. The literature tends to focus on the limitations of search engine indexing. Search engines can do little more than index what publishers make available. Retrieval problems stem from authors' lack of indexing knowledge and limitations in HTML for describing document format. For example, many Web page authors do not utilize meta tags or utilize them inappropriately. Web document descriptions are limited to the URL, title field, hyperlinks and broad attributes such as graphics. These makes content and field indexing difficult. Essentially, few Web-based devices exist for maximizing Precision and Recall. Hyperlinks are one of the few such devices available, but these are generally sporadically assigned. Web publishers should have the responsibility of preparing publications in a way that facilitates retrieval. Understanding basic indexing principles can help. Following this line of argument, let us briefly examine basic indexing principles.

2. Indexing Theory

Essentially, indexing and searching are all about finding the right words. In retrieving an item, something about it should indicate how it relates to the query. The idea is to match the words in a user's mind with those used in a document. Unfortunately, the words don't always mean the same thing. Vocabulary control or the lack of it, creates one of the biggest problems in finding desired documents, or avoiding the retrieval of unwanted documents. This can range from simply not knowing which words to use, to variant meanings associated with a word. The concept-mapping feature of search engines is sufficient to cluster related documents, but do not allow for vocabulary control. The responsibility of providing the "right" words rests with both the searcher and the document author. Indexes provide a critical link between the information need and relevant documents.

When searching for information, users are directed to the index. Most indexes are inverted, meaning that the index displays a term and provides a reference which point back to the source document or documents, which contain that word. Indexes can also contain a variety of syntactic and semantic devices that identify the role of a word in context, or link a subject to other relevant concepts. The purpose of a good index is to retrieve relevant items from a set of related items, and to minimize the retrieval of unimportant items.

3. Indexing Practice

Indexing practices and tools are designed to maximize Precision and Recall. For example, the traditional "see" cross-reference is a Recall device. Link and role indicators are Recall and Precision devices respectively. Other devices include various types of cross-references, scope notes, parenthetical expressions, term weighting and authority control. Thesauri are the principal tools for authority control. Indexes must be constructed to truly reflect content, and to accommodate the mind frame of a wide range of information seekers. The document producer must also share in the responsibility of providing appropriate terms. A document should accurately convey the intended meaning, and provide searchable field tags to discriminate between document components.

There are three main types of indexes. These are:

- Pre-coordinate indexes (classification schemes, such as Library of Congress)
- Post-coordinate indexes (standard computer retrieval)
- Some variant flavor of both (faceted classification, human indexer selection)

Classification schemes reflect the relevance judgment of subject experts in assigning documents to major subject categories. Most classification schemes are hierarchical arrangements, with more specific classes falling under broader ones. Yahoo, Alta Vista and numerous other search engine providers, use this system. The problem with pre-coordinate schemes is that it is difficult to ascertain in advance what prompts a relevance judgment from a user. Thus, the danger in pre-coordinate schemes is that relevant information may not be classified to suit individual needs. Also, coverage of the knowledge universe is limited drastically by the number of human indexers. The obvious answer is to provide an *ad hoc* classification, determined by an individual's information need. This is the basis of post-coordinate schemes. Essentially, if a user's query can be thought of as a classification, all documents meeting the query parameters fall into that class for that given search. Computer searching provides a post-coordinate type of classification. It is made possible by search engines, which index every word in a Web site. Some search engines utilize stop word lists as an attempt at vocabulary control. Most Web search providers use a combination of pre and post coordinate indexing. For the most part, search engine indexing devices favor increasing Recall, but at the expense of lowering Precision. In order to achieve a balance, indexers need to consider basic design criteria.

4. Design Features of Indexes

Four major design criteria govern a good index:

- **A** accuracy
- **C** consistency
- **E** exhaustively, and
- **S** specificity

Good indexes are accurate. They contain no errors, such as typos, blind cross-references or ambiguities. At the very least, it should be easy to ascertain that a subject or concept is present. Accuracy may also reflect the ability to show relationships across word variants, automatically up posting to spelling variations, being mindful of case sensitivity, and being able to distinguish between proper and common terms. Accuracy may also have to do with keeping an index up-to-date.

Consistency refers to the consistent application of indexing rules. In human indexing we talk about inter and intra indexing consistency. That is, does a given indexer always index the same across documents, or would all indexers index a given document the same. In machine indexing, intra-indexing consistency is usually not a

problem although a searcher may have difficulty ascertaining how the indexing actually occurs. Inter indexing across search engines is a real problem, particularly between directory and keyword search engines.
Exhaustivity refers to how completely a document is actually indexed. Some search engines index only portions of a document. Simple frequency counts of term occurrence are not sufficient to justify indexing. The more index terms posted to a document, the higher the Recall. However, exhaustive indexing can lower Precision. Document type and other elements are needed to be truly exhaustive. Exhaustivity also relates to the extent that document are retrieved from a universe.

Specificity involves the level at which an index entry describes document content. If a document deals with cocker spaniels, it should not be indexed under dogs. If the document is also indexed under the subject heading of dogs and related synonyms, links should indicate that the document deals with a particular breed of dog. Semantic differences should be preserved. Essentially, documents should be indexed at the level they present a subject, but should reflect synonymous relationships. Concept indexing practices need to pay more attention to this criterion. Let us turn our attention to things an author can do that facilitate retrieval.

5. Web-based Indexing

To facilitate derivative indexing, document producers should use great care in selecting appropriate terminology to reflect meaning. Documents should begin with informative titles, the use of headings and sub-headings, and structured paragraphs. Documentation should include things like author, date of publication, source, and last revision. These fields are not always searchable in a document, nor is the content amiable to retrieval from different disciplines. Multimedia formats are especially difficult to index. In addition to tagging "free" text document fields, externally derived terms may be added to facilitate retrieval. In modern parlance, this is called "meta data" or meta information".

Meta tags provide a convenient way for document indexing. One author describes meta tags "as attaching a label to an object, such as a can of peas or a package of light bulbs. The label provides information about the contents of the container without actually having to open the container. [http:/www.imsproject.org/metadata] Meta tags are organized into categories, or fields. Each field represents some characteristic of the document or the contents. Meta tags go in the HEAD section of the HTML document, generally after the TITLE tags. They are sometimes the first section to be indexed by a search engine. The two most important Meta tags are those describing the document, and words, which have, clear contextual meaning, such as descriptors derived from a controlled vocabulary. Good indexing tools also guard against malpractice designed to retrieve pages with little relevance to the index terms. Meta tagging lacks the sophistication of well-established knowledge classification schemes. Current schemes tend to reflect practice within a particular industry.

6. XML

Two meta tag schemes currently in use are XML and the Dublin Core. Extensible markup Language (XML) was officially adopted as a standard by the World Wide Web Consortium (W3C) in February of 1998. The W3C calls XML a common syntax for expressing structure in data, or structured authoring. Structured data refers to data that is tagged for its content, meaning, or use. [http://builder.cnet.com/Authoring/Xml20/ss01.html] For example, an XML tag could explicitly identify the type of information <BYLINE>, author of a document <AUTHOR>, cost in an inventory list <PRICE>, all the way down to any level of detail <DOGFOODBRAND>.

XML is a web-based scripting language that promises to provide more efficient applications, and increase both Precision and Recall. XML can specifically categorize data within Web pages, word processing documents, e-mail messages and so on using defined dictionaries of specialized grammar called Document Type Definitions (DTD). Specific DTDs are being developed for various applications. For example, the Microsoft Channel Definition Format (CDF) which describes active channel content, and push vendor Marimba's Open Software Description (OSD) which describes software components. XML can act as a Dewey Decimal System for the Internet. XML overcomes the limitations of HTML by explicitly describing document formats and contents. Essentially, XML

allows Web publishers to insert meta tags into their pages. Browsers can read the field codes but would not display them. XML meta tags can also permit browsers to manipulate data without going back and forth to the server. By separating structure and content from presentation, XML will especially benefit people who produce documents intended to appear across multiple media. Another potential application will be the rebirth of Push. Push refers to developing individualized profiles and sending relevant information to the desktop when it appears on the Web.

7. Dublin Core Background

Until XML matures, one of the few formalized definitions for the use of meta tags is the Dublin Core Metadata Element Set. [(http://purl.org/dc] Essentially, the Dublin Core consists of core meta tags that provides information about document and document-like objects. Document elements, that is the meta tags, are represented by descriptive names intended to convey a common semantic understanding of the element. A tremendous advantage is that the Dublin Core can be implemented in HTML for a wide variety of documents. Thus in formulating a search, the query could contain a reference to the element and the searcher could more accurately determine document specific information. The Dublin Core is an approximation of what libraries have been doing with bibliographic description, such as MARC. Element descriptions have been developed for 15 categories, including: title, author, subject or keywords, description, publisher, other contributors, date, resource type, format, resource identifier, source, language, relation, coverage, and rights. An example of a meta tag for author might look like:
<META NAME="DC.author" CONTENT+"Andrew Torok">

The Dublin Core provides for the concept of links and roles mentioned earlier. Roles can be described using SCHEME qualifiers. In the event there are existing schemes for coding the elements, these are also indicated in the meta tags. An example might by recommended syntax for representing proper names. Schemes would refer to major standards and conventions, such as those emanating from the American National Standards Institute (ANSI) The schemes are referred to by using links to a source such as a URL. The Dublin Core also provides other qualifiers for data description. For example, the SCHEME may refer only to an existing coding system. For specific local information, a TYPE qualifier could be used. A variety of organizations are working on the syntax for the elements, but have not been widely accepted. For example, Educom has created meta tags for educational documents and on-line training courses. [http://www.ott.navy.mil/1_4/adl/educom.htm]

Most current Web authoring tools have no provision for automatically coding for Dublin Core. Thus the meta tags must be keyed in directly. Also, not all search engines recognize the Dublin Core syntax. On the other hand, There are commercial Web page developers who consider the Dublin Core meta tags an integral component of their publications.

8. NIU Dublin Core

The College of Education at Northern Illinois University is experimenting with the Dublin Core set for Web support of a basic education class. The class is a survey of educational practice and pedagogy. Various ASCII and binary coded documents are indexed according to the Dublin core. Publication come from locally generated documents, documents downloaded from the Web, and links to relevant Web sites. Copyright issues notwithstanding, locally generated documents may consist of original material as provided by teachers and students, or various print and electronic publications. A macro using Visual Basic converts Microsoft Word documents and prompts for the Dublin Core Metatags. Publications are made available on the university Web site, and can be accessed from a course Web page.

An ASCI index card is created for binary coded documents. The reflective activities are CGI scripted so students can complete the forms and send them back to the teachers. Access to the documents is made available through a free Sun Microsystems search engine, called SWISH-E, that can recognize the Dublin Core data elements. [HTTP://SUNSITE.BERKELEY.EDU/swish-e/] Actually, SWISH-E can be configured to search various document fields. If desired terms are not found in the meta data, it automatically continues on to find the words in the document body. Index terms are selected from various thesauri and vocabulary lists. Identifiers are also generated

from various existing standards. Graduate assistants are typing in the 15 fields

The population served at any given time consists of several hundred students, across several sections. One advantage of using the Dublin core is to set some uniformity across the various sections of the course. This helps the different instructors present information consistently, provides equitable access to resources, and is advantageous for distance education. The general objective is to provide students access to a core set of publications which reflect course objectives and support collaboration. Despite various instructors contributing numerous print and electronic publications, Precision is quite good. The uniform document and content descriptions afforded by the Dublin meta data permit pinpoint Boolean searching.

One of the biggest problems is that the programming and indexing take time. As the document collection grows, another problem will be that documents reflecting a rather narrow subject area may appear redundant. However, the problem should not be as severe as relying totally on broad Web searches. Non-text publications pose a particularly difficult problem as items are scanned in, downloaded from the Web, or input from a digital camera. At this time, the indexing and retrieval seem to work quite well. As external indexing in the form of XML becomes standard, students may well be able to locate relevant publications from the Web. This will allow for greater collaboration with other colleges. This will also lessen the need for faculty-generated selection and acquisition, and more fully support a constructivist environment. As XML document indexing and retrieval mature, the use of Push technology may become a feasible method for maintaining a specialized Web site. A Push model to support college curricula has been described in the literature. [Torok, 1997]

9. Conclusions

In conclusion, there appears to be continued improvements in Web document indexing and the ability of search engines to reflect user vagaries. Ultimately, the goal should be to maintain the freedom of expression characteristically associated with the Internet and the World Wide Web. There should be equal room for scholarly publications, cloaked in their social and political mantras, and figments of the imagination uninhibited by editorial constraints. The objective language of disciplines should not remain a mystery, nor should it give way to meta language purely to suit the needs of those unwilling to pay the price of formal learning. There does not need to be a watering down of academic disciplines and their curricula. If publishers are held to accurately describing the nature of their publications, users will have greater success in identifying relevant publications, or avoiding undesirable ones. Adhering to good indexing practices, and utilizing new technologies such as XML will help.

10. References

Torok, A., & West, J. (1999) The Impact of Information Literacy on Web-Based learning. http://coe.cedu.niu.edu/~torok/aect01.html/ (last visited July 3, 1999)

http://www.imsproject.org/metadata (visited July 3, 1999)

http://builder.cnet.com/Authoring/Xml20/ss01.html (visited July 3, 1999)

http://purl.org/dc (visited July 3, 1999)

http://www.ott.navy.mil/1_4/adl/educom.htm (last visited July 3, 1999)

http://SUNSITE.BERKELEY.EDU/swish-e/ (last visited July 3, 1999)

Torok, A. (1998). Push or Pull: Internet Support for Education. *XIII Congreso Internacional Sobre Tecnologia Y Educacion a Distancia, 1997,*. Universidad Estatal A Distancia, San Jose, Costa Rica. 644–653.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

REPRODUCTION BASIS

☒ This document is covered by a signed "Reproduction Release (Blanket) form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").